O'REILLY®

Data and Electric Power

From Deterministic Machines to Probabilistic Systems in Traditional Engineering

Sean Patrick Murphy

Strata+ Hadoop
WORLD

Make Data Work

strataconf.com

Presented by O'Reilly and Cloudera, Strata + Hadoop World is where cutting-edge data science and new business fundamentals intersect— and merge.

- Learn business applications of data technologies
- Develop new skills through trainings and in-depth tutorials
- Connect with an international community of thousands who work with data

Job # 15420

Data and Electric Power

From Deterministic Machines to Probabilistic Systems in Traditional Engineering

Sean Patrick Murphy

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Data and Electric Power

by Sean Patrick Murphy

Copyright © 2016 O'Reilly Media, Inc. All rights reserved.

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Shannon Cutt
Production Editor: Nicholas Adams
Interior Designer: David Futato
Cover Designer: Randy Comer
Illustrator: Rebecca Demarest

March 2016: First Edition

Revision History for the First Edition

2016-03-04: First Release

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. *Data and Electric Power*, the cover image, and related trade dress are trademarks of O'Reilly Media, Inc.

While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

978-1-491-95104-0

[LSI]

Table of Contents

Data and Electric Power

Introduction

Energy, manufacturing, transport, petroleum, aerospace, chemical, electronics, computers...the list of industries built by the labors of engineers is substantial. Each of these industries is home to hundreds of companies that reshape the world in which we live. Classical, or traditional engineering itself is built upon a world of knowledge and scientific laws. It is filled with determinism; solvable (explicitly or numerically) equations, or their often linear approximations, describe the fundamental processes that engineers and industries have sought to tame and harness for society's benefit.

As Chief Data Scientist at PingThings, I work hand-in-hand with electric utilities both large and small to bring data science and its associated *mental models* to a traditionally engineering-driven industry. In our work at PingThings, we have seen the original, deterministic models of the electric power industry not getting replaced, but subsumed by a stochastic world filled with increasing uncertainty. Many such industries built by engineering are undergoing this fundamental change—evolving from a deterministic machine to a larger, more unpredictable entity that exists in a world filled with randomness—*a probabilistic system*.

Metamorphosis to a Probabilistic System

There are several key drivers of this metamorphosis. First, the grid has increased in size, and the interconnection of such a large number of devices has created a complex system, which can behave in unforeseeable ways. Second, the electric grid exists in a world filled with stochastic perturbations including wildlife, weather, climate, solar phenomena, and even terrorism. As society's dependence on reliable energy increases, the box that defines the system must be expanded to include these random effects. Finally, the market for energy has changed. It is no longer well approximated by a single monolithic consumer of a unidirectional power flow. Instead, the market has fragmented with some consumers becoming energy producers, with dynamics driven by human behavior, weather, and solar activity.

These challenges and needs compel traditional engineering-based industries to explore and embrace the use of data, with an understanding that not all in the world can be modeled from first principles. As an analogy, consider the human heart. We have a reasonably complete understanding of how the heart works, but nowhere near the same depth of coverage of how and why it fails. Luckily, it doesn't fail often, but when it does, the results can be catastrophic. In healthy children and adults, the heart's behavior is metronomic and there is almost no need to monitor the heart in real time. However, after a coronary bypass surgery, the heart's behavior and response to such trauma is not nearly as predictable; thus, it is monitored 24/7 by professionals at significant but acceptable expense.

To gain even close to the same level of control over a stochastic system, we must instrument it with sensors so that the data collected can help describe its behavior. Quickly changing systems demand faster sensors, higher data rates, and a more watchful eye. As the cost of sensors and analytics continues to drop, continuous monitoring for high-impact, low frequency events will not remain the exception but will become the rule. No longer will society accept such events as unavoidable tragedies; the "Black Swan" catastrophe will become predictably managed and the needle will have been moved. Just ask Paul Houle, a senior high school student in Cape Cod, Massachusetts, how thankful he is that his Apple Watch monitored his pulse during one particular football practice—"my heart rate showed me it was double what it should be. That gave me the push to go and seek help"—and saved his life.

Integrating Data Science into Engineering

Data can create an amazing amount of value both internally and externally for an organization. And data, especially legacy data—

data already collected and stored but often for different reasons—comes with a significant set of costs. In exploring the role of data within the traditional engineering industry, it's essential to understand the *ideological chasm* that exists between engineering based in the physical sciences and the new discipline of data science. Engineers work from first principles and physical laws to solve very particular problems with known parameters, whereas data scientists use data to build statistical and machine learning models and learn from data. In fact, data can *become the models*.

Driving the data revolution has been the open source software movement and the resulting rapid pace of tool development that has ensued. Not only are these enabling tools free as in beer (cost no money to use), they are free as in speech (you can access the source code, modify it, and distribute it as you see fit). As a result, new databases and data processing frameworks are vying for developer *mindshare* as much as for market share. While a complete review of open source software is far beyond the scope of this book, we will examine certain time series databases and platforms, as they relate to the field of engineering. In engineering, numeric data often flows into the system at consistent intervals. Once the data is stored, we need to create some form of value with the data. We will take a quick look at Apache Spark, a popular engine for fast, big data processing, and other real-time big data processing frameworks.

Finally, we will explore a specific problem of national significance that is facing the electric utility industry—the terrestrial impact of solar flares and coronal mass ejections. We'll walk through solutions from the field of traditional engineering, and consider how they contrast with purely data-driven approaches. Finally, we'll examine a hybrid approach that merges ideas and techniques from traditional engineering and data analytics.

While software engineers have also helped to build some of our greatest accomplishments, we will use the term engineer throughout this book in its classical or traditional sense: to refer to someone who studied civil, mechanical, electrical, nuclear, aerospace, fire protection, or even biomedical engineering. This traditional engineer most likely studied physics and chemistry for multiple years in college along with enduring many semesters of calculus, probability, and differential equations. Engineering has endured and solidified to such an extent that members of the profession can take a series of licensing exams to be certified as Professional Engineer. We will not devolve into the debate of whether software engineers are truly engineers. For a great article on the topic and over 1500 comments to read, try this piece from The Atlantic. Instead, remember that for the remainder of this short book, the word engineer will not refer to software engineers or even data engineers, an even more nebulous term.

From Deterministic Cars to Probabilistic Waze

The electric power industry is not the only traditional engineering-based industry in which this transformation is occurring. Many legacy industries will undergo a similar transition now or in the future. In this section, we examine an analogous transformation that is taking place in the automobile industry with the most deterministic of machines: the car.

The inner workings of the internal combustion engine have been understood for over a century. Turn the key in the ignition and spark plugs ignite the air-fuel mixture, bringing the engine to life. To provide feedback to the system operator, a static dashboard of analog or digital gauges shows such scalar values as the distance travelled, current speed in miles per hour, and the revolutions per minute of the engine's crankshaft. The user often cannot choose which data is displayed and significant historical data is not recorded nor accessible. If a component fails or is operating outside of predetermined thresholds, a small indicator light comes on and the operator hopes that it is only a false alarm.

The problem of moving people and goods by road started out relatively simple: how best to move individual cars from point A to point B. There were limited inputs (cars), limited pathways (roads), and limited outputs (destinations). The information that users required for navigation could be divided into two categories based on the rate of change of the underlying data. For structural, slowly evolving information about the best route, drivers used static geographic visualizations hardcoded on paper (i.e., maps) and then translated a single route into hand-written directions for use. On the day of publication however, most maps were already outdated and no longer reflected the exact transportation network. Regardless, many maps languished in glove compartments for years, even though updated versions were released annually.

For local, rapidly changing data about the optimal path—the roads to take and the roads to avoid as a function of time of day and day of week—the end user could only learn via trial and error over numerous trips. This hyper-local knowledge was not disseminated to others—or, if it was, the information was only shared with a select few. Specific road conditions were not known ahead of time, and only broadcast via radio and local news. Thus, local, stochastic perturbances such as sunshine delays,[1] accidents, rubbernecking, and weather conditions could drastically affect drivers and commute times.

Over the last one hundred years, Americans have become more and more dependent on cars and the freedom that they represent. Fast forward to 2015. The car, the deterministic machine and previously the heart of the personal transportation ecosystem, has become a single component in a much larger, stochastic world. To function effectively much closer to the system's capacity limits, society must coordinate hundreds of thousands of vehicles in as efficient a fashion as possible, given complex constraints such as highway structure and geography with numerous random effectors including traffic patterns, work schedules, and weather patterns. The need to drive more efficiency into the current system requires rethinking the problem at a higher level.

> We cannot solve our problems with the same level of thinking that created them.
>
> —Albert Einstein

Fortunately, a significant percentage of cars have been unintentionally instrumented with smartphones: a relatively inexpensive sensor platform equipped not only with GPS and accelerometers but also, and crucially, high bandwidth data connections. At first, smartphone applications like Google Maps offered digital versions of static maps with one key element of feedback: a blinking blue dot showing the driver's location in real-time. As Google leveraged historical trip data, Google Maps could provide more optimal paths for its users.

Waze extended this idea further and built a community of users who were willing to provide meaningful feedback about current road

[1] Traffic delays, usually for west- or east-bound drivers, caused when the sun is low in the sky and impairs driver vision, forcing cars to slow down

conditions. The Waze platform then broadcasts this information back to all app users to provide alternative route options dynamically and tackle the problem of stochastic perturbations to traffic patterns. The next step in these products' evolution is to suggest different paths to different drivers attempting to make similar trips, thus spreading traffic across the existing roadways to relieve congestion, and more effectively use the existing infrastructure. Although the drivers are still in control of their cars, data-driven algorithms are providing feedback in real time.

These advancements would not be possible without the existence of numerous enabling technologies and data systems built completely independently of the transportation system. One such data system, the Global Positioning System, was first conceived of by two physicists at the Johns Hopkins University Applied Physics Laboratory monitoring the Sputnik 1 satellite in 1957.[2] Today, a constellation of 32 satellites in six approximately circular orbits continuously stream real-time location and clock data to ground-based receivers that can use this data to compute location anywhere on Earth, assuming at least 4 satellites are in view.

On the hardware side, Moore's Law[3] has helped make personal, portable supercomputers a reality complete with miniaturized sensor systems. On the side of software infrastructure, we have watched the rise to dominance of virtualized infrastructure as a service (IaaS), platforms as a service (PaaS), and software as a service (SaaS). Whether you want to build a large scale computing platform from scratch using virtual instances from an IaaS such as Amazon Web Service, Google Compute Engine, or Microsoft Azure, or simply use someone else's machine learning algorithms as a service from a PaaS such as IBM's Watson Analytics, you can. What was once a massive, upfront capital expense has transformed into an on-demand fee, proportional to what is consumed. As these capabilities have evolved, so too has the data science software stack. All of these factors have enabled services such as Waze to arise and begin to transform the more than a century old automobile industry from

[2] Klingaman, W. K. (1993). APL, fifty years of service to the nation: A history of the John Hopkins University Applied Physics Laboratory. Laurel, MD: The Laboratory.

[3] Moore's Law is the observation by the former CEO of Intel, Gordon Moore, that the number of transistors in a microprocessor tended to double every two years.

what started as a small number of deterministic machines to a complex, probabilistic system.

A Deterministic Grid

> In mathematics and physics, a **deterministic** system is a system in which no randomness is involved in the development of future states of the system. A **deterministic** model will thus always produce the same output from a given starting condition or initial state.
>
> —Wikipedia

The delivery of electric power has become synonymous with utility; plug an appliance into the wall and the electricity is just there. The expectation of *always on, always available* has permeated the consumer psyche from telephone, power, and more recently Internet connectivity. Electrification even earned the distinction as the greatest engineering achievement of the 20th century from the National Academy of Engineering. What has enabled this feat of predictability are the laws of physics discovered in the preceding centuries.[4]

In 1827, Georg Ohm published the now famous law that bears his name and states: “the current across a conductor is directly proportional to the applied voltage. Thus, a voltage applied to a power line with known characteristics will result in a computable current flow.” In the 1860s, James Clark Maxwell laid down a set of partial differential equations that formed the basis for classical electrodynamics and ultimately, circuit theory. These equations describe how electric currents and magnetic fields interact and underlie contemporary electrical and communications engineering, and are shown both in differential and integral form in Table 1-1.

Table 1-1. Point and Integral forms of Maxwell’s Equations. Variables in bold font are vectors. E is the electric field, B is the magnetic field, J is the electric current, and D is the electric flux density.

Name	Differential Form	Integral Form
Ampere’s Circuit Law	$\nabla \times \mathrm{H} = \mathrm{J}_c + \frac{\partial \mathrm{D}}{\partial t}$	$\oint \mathrm{H} \cdot dl = \int_S \left(\mathrm{J}_c + \frac{\partial \mathrm{D}}{\partial t} \right) \cdot d\mathrm{S}$
Faraday’s Law of Induction	$\nabla \times \mathrm{E} = -\frac{\partial \mathrm{B}}{\partial t}$	$\oint \mathrm{E} \cdot dl = \int_S \left(-\frac{\partial \mathrm{B}}{\partial t} \right) \cdot d\mathrm{S}$

[4] Greatest Engineering Achievements of the 20th Century, National Academy of Engineering

Name	Differential Form	Integral Form
Gauss's Law	$\nabla \cdot D = \rho$	$\oint_S D \cdot dS = \int_V \rho dv$
Gauss's Law for Magnetism	$\nabla \cdot B = 0$	$\oint_S B \cdot dS = 0$

These laws and many others, such as Kirchoff's laws, enabled models of real and complex systems, like the power grid, to be built from first principles, describing how something works from immutable laws of the universe. With these models, one can arguably say that they completely understand the system. That is, given a set of conditions, important system values can be determined for any time either in the past or the future. Of course, this understanding is constrained by the set of assumptions under which those equations hold true.

Moving Toward a Stochastic System

> Stochastic is synonymous with "random." The word is of Greek origin and means "pertaining to chance" (Parzen 1962, p. 7). It is used to indicate that a particular subject is seen from point of view of randomness. Stochastic is often used as counterpart of the word "deterministic" which means that random phenomena are not involved. Therefore, stochastic models are based on random trials, while deterministic models always produce the same output for a given starting condition.
>
> —Vincenzo Origlio[5]

The electric grid, which started as a *deterministic machine* based on a model of one-way power flow from large generators to customers and governed fundamentally by well-known and understood mathematical equations, has transformed into a *probabilistic system*.

We see three key drivers of this metamorphosis:

1. Though many of the deterministic components, such as generators and transformers, have well-described mechanistic models, or operate in regions sufficiently approximated by linear relationships, the interconnection of so many devices has created a complex system. While a critic may argue that the uncertainty arising from a complex system differs from a truly random

[5] Origlio, Vincenzo. "Stochastic." From MathWorld—A Wolfram Web Resource, created by Eric W. Weisstein.

model, the outcome is similar—we aren't sure what happens for a given set of initial conditions. Adding to this technical complexity is one of business complexity. Many of the once vertically integrated utilities have been transformed, with separate companies taking ownership and responsibility for the power plants, transmission and delivery, and even marketing to the end consumers.

2. The grid exists in a world filled with what were once considered external random challenges to the system. Such stochastic phenomena as bird streamers, galloping lines, geomagnetic disturbances, and vegetation overgrowth have plagued system operators for decades. As the demands placed on the grid increase and the system operates closer to the edge of its capacity, these random effects must now be considered part of the greater system as a whole.
3. The market for energy has fragmented. It has transitioned from a simple market, well approximated by a monolithic consumer of a unidirectional power flow, to a fragmented, multidirectional market of individual consumers and producers, where consumption and production is driven by truly random phenomena, such as weather and solar activity.

On top of these three sources of stochasticity, society's reliance on electricity has never been greater. The loss of electricity can translate to billions of dollars of damage and lost opportunity in only a few days.[6] Reliable electricity is required by every industry and every person in the industrialized world, so much so that lives and national security depend on its availability every second of every day. As a result, the national power grid must directly address these new challenges and evolve from a deterministic machine to a probabilistic grid.

Stochastic Perturbances to the Grid

The nation's electric grid stretches over all 50 states via 360,000 miles of transmission lines (180,000 of those are high-voltage lines), and over 6,000 power plants that exist in dozens of different climates and

[6] J.R. Minkel, "The 2003 Northeast Blackout--Five Years Later," *Scientific American Online*, August 13, 2008.

environments.[7] With such exposure and expanse, the nation's grid faces numerous perturbances from random actors, such as wildlife, weather, space weather, and even humans via cyberterrorism and physical attacks.

Wildlife

The behavior of wildlife of all sizes impacts the grid. Around the turn of the century, Southern California Edison faced a problem of unexplained short circuits in their newest high voltage power lines, some of the highest voltages that had been built to that point (over 200,000 volts).[8]

Eagles and hawks would use the high vantage point that the new power lines provided to spot potential prey. When taking flight from the lines, the birds would relieve themselves of excess mass, creating arcs of highly conductive fluid known as "bird streamers." If this waste was jettisoned close enough to the transmission tower, the streamer served as a low impedance path from the energized line to the metal tower, circumventing the insulators and providing a pathway to ground. This resulted in a short circuit, and subsequently caused the organic material to flashover, completely destroying evidence of the problem's origin. Unsurprisingly, "bird streamers" had not been accounted for in the original design and the resulting short circuits caused brief but mysterious power interruptions every few days.

While bird streamers are no longer a critical infrastructure problem, squirrels still manage to wreak a considerable amount of havoc on the power grid, as do other wildlife. Although precise numbers are impossible to come by, it is estimated that 12% of all power outages are caused by wildlife.

Weather

As everyone has probably experienced, weather of all types can cause disruptions to power delivery. High winds can knock over trees that then take down power lines or even knock over the power

[7] Large Power Transformers and the U.S. Electric Grid, United States Department of Energy, 2012, page 5.

[8] Charles Choi, "The Forgotten History of How Bird Poop Cripples Power Lines," *IEEE Spectrum*, June 10, 2015.

lines themselves. Snow and ice can accumulate on power lines, causing them to sag, increasing resistance to the flow of electricity and potentially causing them to snap.

Less well known is the phenomenon of galloping lines. For lines to "gallop," a number of environmental factors must cooccur. When the temperature drops sufficiently, ice can form on transmission lines in such a fashion as to create an aerodynamic shape. When the wind blows across the line at the correct angle and with sufficient speed, lift is generated on the cable. Since the line is fixed at both ends to a tower or pole, standing waves can be generated, much like a guitar string but of visible amplitude. If the wind is strong enough, the standing waves can be of sufficient amplitude and force to tear the line from the tower. This behavior is best seen in a video.

Space weather

Until now, the random disturbances discussed affect localized sections of the power grid, usually on the distribution side of the grid. pace weather changes that.[9] On March 13, 1989, a severe geomagnetic storm caused a nine-hour blackout in Quebec.[10] In 1859, the so-called Carrington Event occurred; a large solar flare caused telegraphs to work while disconnected from any power source and the aurora borealis to be seen as far south as the Caribbean.[11] If a Carrington-level event happened today, the results would be catastrophic. It takes two years to replace some of the largest transformers in the United States that are instrumental to the grid's operation and could be damaged or destroyed in a large geomagnetic storm. In fact, the threat is severe enough for the White House's National Science and Technology Council to publish a National Space Weather Action Plan in October 2015:

> Space-weather events are naturally occurring phenomena that have the potential to disrupt electric power systems; satellite, aircraft, and spacecraft operations; telecommunications; position, navigation, and timing services; and other technologies and infrastruc-

[9] NERC, 2012 Special Reliability Assessment Interim Report: Effects of Geomagnetic Disturbances on the Bulk Power System, February 2012.

[10] James L. Green, Scott Boardsen, Sten Odenwald, John Humble, Katherine A. Pazamickas, "Eyewitness reports of the great auroral storm of 1859," *Advances in Space Research*, Volume 28, Issue 2, 2006.

[11] *Ibid*

> tures that contribute to the Nation's security and economic vitality. These critical infrastructures make up a diverse, complex, interdependent system of systems in which a failure of one could cascade to another. Given the importance of reliable electric power and space-based assets, it is essential that the United States has the ability to protect, mitigate, respond to, and recover from the potentially devastating effects of space weather.

We will go deeper into this threat later in the book.

Cyber attacks and terrorism acts

Intentional actions, either electronically or via physical action, are a very real and unpredictable threat to the power grid. In what is the first acknowledged example, a cyber attack using the BlackEnergy Trojan on a regional Ukrainian control center left thousands of people without power at the end of December in 2015. More famously, the Stuxnet computer worm, developed by the US, damaged multiple centrifuge machines used to enrich Uranium in Iranian nuclear facilities in 2010. The Stuxnet worm itself was a sophisticated piece of software, attacking a very specific layer of the Supervisory Control And Data Acquisition (SCADA) systems software written by Siemens, running on computers not directly connected to the Internet.[12] While there are no publicly known, successful cyber attacks on the US grid, one must assume that there will be in the future.

Cyber attacks are not the only concern for our nation's power infrastructure. While the following might read like the first chapter of a Tom Clancy novel, the sniper attack on the Metcalf Transmission Substation outside of San Jose, California was all too real. Shortly before 1 a.m. on April 16th, 2013, fiber optic communications cables were cut south of San Jose. Several minutes later, another bundle of cables near the Metcalf Power Substation was also cut. Over the next hour, multiple gunmen opened fire on the substation, targeting oil tanks critical to cooling the transformers. By 1:45 a.m., the attack was complete. More than one hundred 7.62x39mm cartridges were found on site, all wiped clean of fingerprints. Over 52,000 gallons of oil had leaked out resulting in overheating and damage to seventeen transformers, requiring weeks to repair at a cost of over $15 million

[12] S. Karnouskos, "Stuxnet Worm Impact on Industrial Cyber-Physical System Security." *37th Annual Conference of the IEEE Industrial Electronics Society* (IECON 2011), Melbourne, Australia, 7-10 Nov 2011. Retrieved 20 Apr 2014.

dollars. All evidence points to a well-prepared and professional attack. Given the fact that the power grid stretches over vast portions of the continent, it is simply not possible to cost effectively guard such a large physical footprint.[13,14]

Probabilistic Demand

The electric industry was considered a natural monopoly and was operated as such for many decades. Power generation, transmission, and distribution were all controlled by large, vertically-integrated utilities. Under this model, the marketplace for electricity was practically monolithic. One way of thinking about the current power grid is like a volcano. Each day, the volcano erupts (a certain amount of power is generated per day based on predictions from the previous day) and the lava flows down the mountainside. Similarly, power flows through the transmission and then distribution portions of the grid, to the end residential or commercial consumer. If too much power is generated, there is no way to store it, so it is wasted. If too little power is generated, either more power must be made available or brownouts—dimming of the lights reflecting a voltage sag and effort to reduce load—or even blackouts can occur.

Due to the deregulation of the electric industry in many parts of the country, the market has changed dramatically and become open to a large number of new variables. Even so, this market structure was simple enough to be effectively modeled using a deterministic approach. Variables such as day-ahead demand, the timing of peak demand, available generation, and fuel availability could be accurately estimated.

Today the world is much more complicated, and estimating those same variables has become difficult. In the words of Lisa Wood, Vice President of The Edison Foundation, and Executive Director at the Institute for Electric Innovation:

> No longer an industry of one-way power flows from large generators to customers, the model is beginning to evolve to a much more distributed network with multiple sources of generation, both large

[13] Richard A. Serrano, Evan Halper, "Sophisticated but low-tech power grid attack baffles authorities," *Los Angeles Times*, February 11, 2014.

[14] Alexis C. Madrigal "Snipers Coordinated an Attack on the Power Grid, but Why?" *The Atlantic*, February 5, 2014.

and small, and multidirectional power and information flows. This is not a hypothetical future. It's already unfolding.

Solar panels

The traditional "volcano" model of energy consumption is being disrupted in numerous ways that are all functions of random variables. Homeowners are installing solar panels on their roofs. At the right latitude and environment, these panels can supply more energy than the homeowner needs and actually return energy to the grid. As a result, an estimated 1 million households could become energy producers by 2017 (there are approximately 125 million households in the US in 2016), decreasing demand on traditional utilities in a very random fashion, dependent on weather and cloud formations.[15] Further stochasticity exists in the adoption of these new renewable energy technologies, as some states are more receptive than others in terms of the applicable regulations and policies.

Home energy storage

Consumer home energy storage systems such as the not-yet-released Tesla Powerwall promise to complement this burgeoning photovoltaic market. While home energy storage helps to smooth out the cyclical and stochastic power generating capabilities of solar and wind energy, it potentially adds more complexity and another element of human behavior to the grid. Even for homes without local energy generation, consumers with home energy storage could purchase energy during times when prices are cheaper and store it for later use.

The electric car

Further adding randomness to the market for electricity is the electric car. The Nissan Leaf has sold over 200,000 units globally as of the end of 2015. Tesla's second car, the model S, has globally sold over 107,000 units as of the end of 2015. As the costs for these models drops and the range of their batteries gets longer, it is likely that sales will only increase. Charging schedules for electric cars add a further large and unpredictable element to the marketplace as they are complex functions of vehicle usage.

[15] Rhone Resch, "Solar Capacity in the U.S. Enough to Power 4 Million Homes," EcoWatch, April 22, 2015.

Wind- and solar-farms

Even larger scale, utility-owned wind- and solar-farms introduce significant randomness into what was once a much more deterministic load on the power grid. In simple terms, a power plant needs to burn a known amount of coal to generate a specific amount of power. However, the production output of a wind-farm and a solar-farm varies unpredictably with the weather. Further, these new renewable sources often do not come online where load growth has occurred. This adds stresses and strains to the transmission and distribution systems, pushing it into operating regimes where it can become more vulnerable to other random phenomena.

Instead of a small number of market participants, there are now a large number of players. Instead of unidirectional energy flow on the distribution system, distributed generators are creating bidirectional flows of energy. The number of consumers is increasing, and the variability amongst consumer behavior is also increasing. Weather impacts generation more so than ever, all while the weather is becoming increasingly unpredictable. The summation of these forces results in a system that is becoming increasingly probabilistic in nature.

Traditional Engineering versus Data Science

Verticals such as the power utilities, chemical production, pharmaceuticals, aerospace, automotive, and most manufacturing companies are only made possible by the hard work of traditional engineers. Yes, oftentimes software programmers (or dare I say software engineers) are involved as well, but we are still using engineer in its traditional sense. Think Scotty from Star Trek, not Neo from The Matrix!

To better understand the difficulties evolving from a traditional engineering industry to one that is data-driven, we will look at what classical engineering is, and how many of these defining characteristics directly conflict with data science and the machine learning revolution.

What Is Engineering?

If you are an engineer, does the following curriculum sound familiar? In your first year, you spend your time studying various mathematics such as geometry and trigonometry and the physical and chemical sciences. In your second and third year, you continue to strengthen your background in mathematics but also learn structural and mechanical engineering, transitioning from the theoretical to the applied. In your fourth year, you might find yourself specializing further and working on a real world project in the field.

Interestingly, this is the engineering curriculum of the École Polytechnique in France, *at the beginning of the 19th century.*[16]

Look across different definitions of engineering and you start to see a pattern. John A. Robins at York University captures this semantic average as five characteristics, starting with the core definition that: "*[e]ngineering is applying scientific knowledge and mathematical analysis to the solution of practical problems.*" He notes that engineers often design and build artifacts, and that these objects or structures in the real world are good, if not ideal, solutions to well-defined problems. Most crucially, engineering "*applies well-established principles and methods, adapts existing solutions, and uses proven components and tools.*"[17]

Fundamental to engineering is the set of underlying models (or conceptual understanding) that describe how a particular part of the world works. Take for example, *electrical engineering*. Ohm's law tells us that the potential difference across a resistor is equal to the product of the current flow and the resistance that the resistor offers. These physical laws and models help the engineer to represent, understand, and predict the world in which he or she works. Most of these laws are approximations, or are only valid given a set of assumptions of which the good engineer is aware. These models, and the ability to predict the behavior of these models, allow the engineer to build solutions to specific problems with known specifications.

On top of these fundamental models, an engineer assembles one or more solutions to a problem. It isn't chance that the word *engineering* is derived from the Latin *ingenium*, which means "cleverness," but this attribute of an engineer is dependent on the ability to accu-

[16] Artz, Frederick B. The Development of Technical Education in France: 1500-1850. Cambridge (Massachusetts): M. I. T., 1966. Print.

[17] John A. Robinson, "Engineering Thinking and Rhetoric"

rately predict how things will work and behave. This, in turn, is derived from the models of how the world works. Thus, the engineer is constrained by the limits of this previously discovered knowledge, and the gaps or cracks between adjacent fields. Her intent is not to discover new knowledge or undiscovered principles, but to apply and leverage scientific knowledge and mathematical techniques that already exist.

A list of the original seven engineering societies in the American Engineers' Council for Professional Development circa 1932 highlight the major branches of engineering: civil, mining and metallurgical, mechanical, electrical, and chemical engineering. These engineering fields were all built on top of previously established scientific knowledge and best practices. Over time, the list of acknowledged engineering disciplines has grown substantially—manufacturing engineering, acoustical engineering, computer, agricultural, biosystems, and nuclear engineering to name a few—but the prerequisite scientific knowledge always came first and laid the foundation for the engineering discipline.

What Is Data Science?

Entire books have been written about what exactly qualifies as data science. Some even incorrectly believe it to be a "flashier" version of statistics. Instead of tackling this amorphous question, we will take a more concrete approach and look at the practitioners of this new field, the data scientist.

Anecdotally, the term "data scientist" was first coined by DJ Patil and Jeff Hammerbacher, when trying to provide human resources with the right label for the job posting that they needed filled at LinkedIn.[18] Drew Conway elegantly visualized the skill sets of this new data scientist in his now infamous but apropos Venn diagram (Figure 1-1); a data scientist was the strange collection of hacking skills, mathematical prowess, and subject matter expertise. While others have added communication as a fourth circle or suggested similar changes, this diagram still does an admirable job of summing up a data scientist.

[18] Anecdote related by DJ Patil at Meetup.com Event in Washington DC, October 10, 2015.

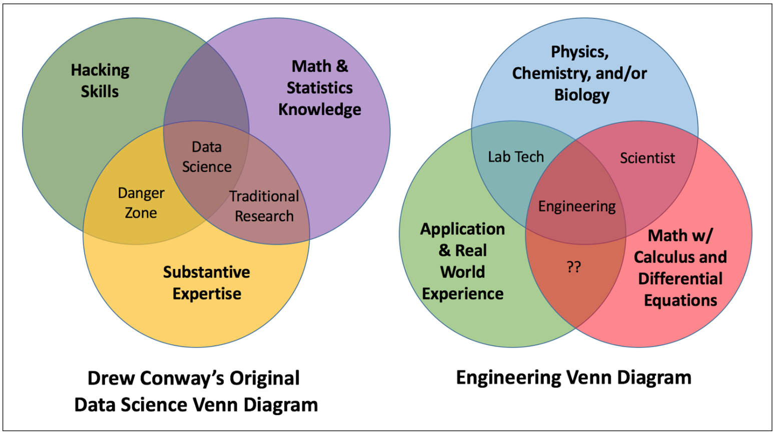
Hacking Skills
Math & Statistics Knowledge
Data Science
Danger Zone
Traditional Research
Substantive Expertise
Drew Conway's Original Data Science Venn Diagram
Physics, Chemistry, and/or Biology
Lab Tech
Scientist
Engineering
Application & Real World Experience
??
Math w/ Calculus and Differential Equations
Engineering Venn Diagram

Figure 1-1. Drew Conway's original data science Venn diagram and what a general engineering Venn diagram might look like

In 2012, Josh Wills tweeted his personal definition; "Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician." All joking aside, this definition perfectly captures the original zeitgeist of the data scientist—an inquisitive jack-of-all-trades whose computer skills are good enough to write usable code and interface with large scale data systems, and with sufficient mathematical chops to understand, use, and even refine statistical and machine learning techniques.

As data science arose out of industry, it is not an abstract subject but an applied one. To ask the right questions and interrogate data intelligently, the practitioner needs to have some depth of knowledge in the relevant field. Once answers are found, the results and their implications must be relayed to individuals who often have no technical background or mathematical literacy. Thus, communication and, even more, storytelling—the ability to construct a compelling narrative around the results of an analysis and the implications for the organization—are key for the data scientist.

Why Are These Two at Odds?

At first glance, traditional engineering and data science seem similar. Engineers, just like data scientists, are often well trained in math. The data scientist is more heavily focused on statistics and probability, while engineers spend more time modeling the physical world with calculus and differential equations. Computers are a tool required by both professions, but the required level of proficiency is quite different. Most engineers have at least some programming experience, but it is often using Matlab. (Don't worry, we won't go off on a rant about how and why Matlab is evil and facilitates the adoption of all kinds of terrible programming habits.) Suffice it to say that scripting solutions to problem sets in Matlab differs from developing production-quality software systems. By definition, data scientists live and breathe data. As this data only lives in the virtual world, strong programming skills are a must. Engineers tend to be users of software tools, whereas many data scientists are creators of software tools and systems.

Engineers have deep subject matter expertise in a particular science, often physics or potentially chemistry or biology. While data scientists also tend to have deep expertise, it can be in a seemingly tangential field, such as political science or linguistics. Further, the engineer's scientific background supplies the models and approximations detailing how the world works. In contrast, the data scientist's subject matter expertise is almost an outlet or representation of her or his intellectual curiosity. Understanding a subject deeply means one is better equipped to formulate more piercing questions during an inquiry into the same or even a different topic. Demonstrating deep knowledge of one area is also a strong indicator that one can achieve a deep understanding of another field.

The engineer supplements her or his foundational scientific knowledge with detailed applied knowledge in the chosen field. For electrical engineers, this could be communications theory or circuits or power systems. These subjects build upon the scientific foundation, applying the principles of physics to solve applied technical challenges. Engineers master these approaches and learn the underlying patterns to then tackle similar problems in the real world; this approach can be considered more deductive in nature. For data scientists, the approach most often used is more *inductive* in nature. Observations as manifested through data can lead to patterns and hypotheses, and then ultimately, to learning about the system under examination. While many will argue whether data science is truly a science, there is often a strong exploratory nature to data-oriented projects.

Diving into this key difference a step further, one of the enabling technologies behind the data science revolution is machine learning, the field "concerned with the question of how to construct computer programs that *automatically improve with experience*."[19] For a much more extensive definition, I recommend the following blog post. Machine learning is a monumental paradigm shift. With the algorithms that have been and are being developed, data is being used to program machines. *Instead of people implementing models and simulations in software, data is teaching computers.*

The Data Is the Model

It might be easier to offer up a simple example to compare and contrast traditional engineering and data science. Take the solved problem of determining the area of a circle. The engineering solution would come from the existing knowledge of a deterministic model that computes the area using two parameters, the constant π and the radius of the circle. This formula works every time. Now, assume for a moment that this compact representation did not exist or was unknown. How else could the area be measured?

One data-oriented technique would be to employ a Monte Carlo simulation. A circle is inscribed in a square of known area and a set of test points [x, y] is randomly distributed throughout the defined system. Each test point is examined for whether the point is within the circle and the result, a yes or no, is recorded. The ratio of the points that fall within the circle to the total points tested multiplied by the area of the square yields the area of the circle. As the number of random points generated and tested increases, an increasingly accurate representation of the circle's area is developed. More data results in a more accurate model. In fact, the data literally becomes the model, as visibly demonstrated in the panels of Figure 1-2.

[19] Mitchell, Tom M. Machine Learning. New York: McGraw-Hill, 1997. Print.

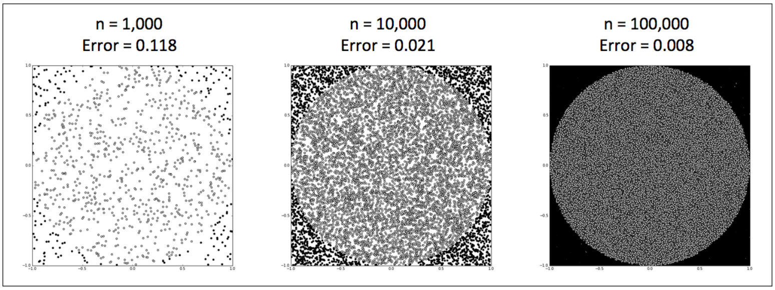

Figure 1-2. Visualization of the results of the Monte Carlo simulation used to find the area of a circle. Points outside the circle are filled black and points inside the circle are left white. Moving from left to right, the number of random test points increases by two orders of magnitude while the error on the estimate of the area decreases.

Extending this example further, we can then take this collection of points, both inside and outside of the circle, and build a classifier from the data, to determine if new points that are added to our system are inside or outside of the circle. The data has now been used to program the machine to compute the area of a circle.

Understanding Data and the Engineering Organization

Data has a rich history of use within traditional engineering organizations. As early as 1928, technicians from electric utilities would drive out to a remote transformer, pull a sample of oil used to cool and insulate the transformer, and then measure the dissolved gases in the oil. The presence of specific gases in the oil is an indirect indicator of the health of the transformer. Fast-forward to the 1970s and the implementation of Supervisory Control and Data Acquisition (SCADA) systems throughout the grid. For the first time, field equipment was monitored continuously and different metrics such as voltage and current flows were recorded every few seconds to give approximate real-time monitoring and control.

Engineers themselves have always valued data; measurements in the lab or the field help to determine if a component or system is working as expected or, even more fundamentally, to verify if their models of the world actually reflect reality. So what has changed? Why talk about data with regards to traditional engineering organizations? Why now?

The answer is simple: money. The cost-benefit analysis for data ossified years ago for many companies. However, while traditional engineering organizations continued to produce relatively consistent electric power or manufacture cars or create plastic compounds, the web companies of the world went ahead and simultaneously reduced the cost of data radically and demonstrated how to increase its value exponentially. *Data has changed from a necessary evil for certain critical tasks to a way of life* and H. James Harrington's quote in 1999 seems almost prescient:

> Measurement is the first step that leads to control and eventually to improvement. If you can't measure something, you can't understand it. If you can't understand it, you can't control it. If you can't control it, you can't improve it.

What product or process doesn't a company want to improve?

Data has become an asset. As measurements are really just data, collecting measurements is now comparable to accruing assets. Eventually, with enough measurements, you get into "big data" territory. Much has been made about this term and the three V's of big data: Volume, Velocity, and Variety.[20] While "big" data is often large in size (although size is always relative to the available tool set, resources, and staffing), this notion is somewhat of a red herring. The "big" in big data actually means important, as in a big deal, and the fourth "V" should be "Value." Scientists have long known that data could create new knowledge via the application of the scientific method. Now, the rest of the world, including management and even government, has realized that data can create value, be it financial, environmental, and/or social value.

The Value of Data

To help understand the value that can be created with data, we take a quick look at Gartner's "analytic value escalator," which first appeared in 2012. The original plot shows the value offered by various categories of "analytics" as a linear function of difficulty, which is really just a surrogate for cost. Further, they group analytics by the time period that they are trying to inform: the present, the past, and

[20] Volume refers to the amount of data being generated. Velocity refers to the rate of generation of the data and Variety to the fact that the data being created ranges from stock values, to Tweets and 4K video.

the future. While I like what Gartner was trying to do, I don't think they got it quite right. An "improved" version shows the relationship between value and cost (or difficulty) for data products not as a linear function but as a saturating exponential (see Figure 1-3).

Note that we switched from "analytics" to full blown "data products." If you are unfamiliar with the term, data products are simply products that are built from data, often but not always involving some form of statistical analysis or machine learning technique. Mike Loukides puts it best: "[a] data application acquires its value from the data itself, and creates more data as a result. It's not just an application with data; it's a data product. Data science enables the creation of data products." Netflix's movie recommendations for its users are one very well known *data product*, but many others exist across industries. As we seek to create value from our data in a reproducible and scalable fashion, we will be talking about data products.

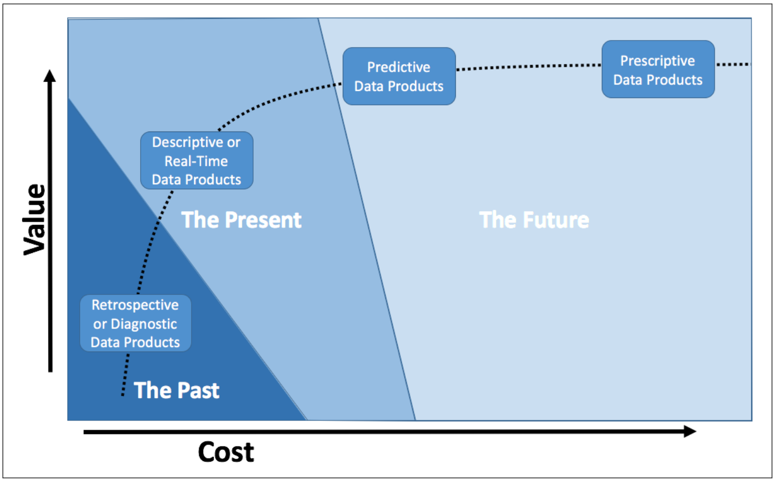
Predictive Data Products
Prescriptive Data Products
Descriptive or Real-Time Data Products
The Present
The Future
Value
Retrospective or Diagnostic Data Products
The Past
Cost

Figure 1-3. The value created plotted against the cost for data products focused on different time periods

The past

We start with understanding the past, answering the questions of what has happened and potentially even why it happened. Why start here? Because, from a technical perspective, this is potentially the easiest timeframe to address, as we have already collected the data and (hopefully) stored it someplace accessible. Borrowing terminology from the next chapter, our data set is bounded or finite, and we can use batch processing software and systems for the analysis (anything from R to Hadoop).

This retrospective or historical data analysis can help us diagnose our current situation, be it a faulty sensor, a faulty transformer, or a poorly performing branch of the organization. We can verify that our understanding of our world is correct, or surface previously unknown relationships.

The present

The next lowest hanging fruit to pluck for data is to describe *the present*, answering the seemingly basic question of what is currently happening. You can think of this as the dashboard in your car or the control room of the electric utility. Often this requires large volumes of data be compressed and summarized for human consumption, a not insignificant task. Further, understanding the present is not a one-time task, but one that must be repeated with the newest data available.

Amazingly, the current state of large swaths of the organization are often unknown and simply glide along in a state of inertial bliss (your heart just keeps on beating, until it doesn't). Here is where organizations can unlock significant value with concerted effort. Even more importantly, monitoring exactly what you have reinforces the existing corporate structure, often a safe path from the traditional career perspective.

The future

Now that we understand the past and where we are, what about the future? Data and predictive models can help us answer this question as well, and seeing into the future can be incredibly useful, with the utility limited only by the accuracy of the predictions. Building predictive models can be challenging, and understanding the limitations of said models even more so. If data scientists were truly capable of perfectly seeing the future, we would all be retired on islands drinking frozen beverages.

Prescriptive analytics are the final frontier of value per Gartner; if we can predict multiple possible futures with data, how do we take action to make a particular one occur? While this potential seems incredibly appealing and valuable (it is at the highest point on the curve), it is also the hardest to realize. The fundamental problem is that organizational structures were simply not designed to predict the future and act on this information. Thus, there are likely many significant structural obstacles blocking the true potential here.

Data source origins

We can break down data sources into two categories; there is data internal to the organization and then data external to the organization. For data internal to the organization, there is data which is already being measured and captured as part of routine operations (or not so routine operations as the case may be). Representing a far larger volume, there is data internal to the organization that is not being captured nor even measured. Going even further afield is all of the relevant data that originates external to the organization.

Why do we care? Thinking about the origins of data sources helps us to understand the value that can be created from them. Shivon Zilis at Bloomberg Beta posted a description of eight classifications of startups whose businesses were built on machine learning. Three of these categorizations relate directly to the topic here and vary based on the source of the data, offering to us insight how larger companies could build valuable data products. First, the *panopticon*, as she puts it, collects a broad data set to tackle problems and questions whose answers have eluded the company or industry. Usually, the data set in question is being collected from multiple sources and multiple stakeholders, crossing internal fiefdoms and organizational boundaries. In the utility space, Genscape serves as an example panopticon. Genscape sets up antennas to monitor the electromagnetic field near high voltage power lines across numerous disparate utilities. Through these measurements, Genscape can infer the amount of power flowing across wide portions of the grid, information that is incredibly valuable to financial traders.

A *laser* collects and leverages a very deep data set for a very focused problem, often with an immediate payoff. The solution is not simply the data but is the unique combination of the data, the algorithms, and the interface provided to the end user. FlightCaster was an excellent example of a laser. Before they were acquired, FlightCaster predicted potential flight delays for travelers based on historical travel data. Finally, gateways create value from data that was too unwieldy to handle before such as video or high-resolution imagery.

An example of this would be to use drone-based aerial footage to quantify and potentially even predict foliage impact on power lines.

The cost of data

Balancing the value that can be created from data is the cost associated with doing so. How much does it cost to acquire the data? Is it being thrown off as "data exhaust" from standard operating procedures? Or, would new processes have to be developed and deployed or new sensor systems developed? And—once the data exists, how expensive is it to store? Does government or organizational policy require security implementation and auditing?

Even if data gets stored, is it only getting archived because regulations mandated the action? And—more importantly—has anyone ever looked at the data to ensure that the archival process was successful? This point might seem obvious but I have seen more than one multi-billion dollar company in a heavily regulated industry spend small fortunes instrumenting devices and archiving data, only to later learn that the data saved was garbage. As a rule of thumb, if no one is consuming the data in your database, you have no guarantees that the data is valid or that it even exists.

Finally, there is the cost of doing something with your data—including the cost of hiring the talent who will perform the work. Newton's First law of motion has an unexpected corollary in the data world—data at rest tends to stay at rest (i.e., unused), unless an external organizational force is applied.

Many have heard that 80–99% of the time spent working on data-related projects is consumed by data wrangling or data munging—acquiring the data, cleaning the data, and then transforming it into a form that is usable for repeated analysis. Organizations that can streamline or automate these processes will reap massive rewards.

Legacy data

Legacy data—that data previously collected for some other purpose —deserves special attention. If data is the new oil, then you might imagine that much of the reserves for older companies are locked away in deep ocean wells or on federally protected land. Data collected 10, 20, 30, or 40 years ago was collected in a vastly different IT environment than today's world of open source software and RESTful APIs. In the past, some data may have been captured on paper, and a determination may have to be made as to the cost effectiveness of digital conversion.

Data captured digitally was often done within closed, commercial software from a third party—through a vendor that may no longer exist. Such third-party companies were typically built based on their ability to make measurements in industrial or commercial settings, and developed software around those needs. Whether for performance or vendor lock-in reasons, the software would often store data only in a *proprietary*, binary format. In best case scenarios, third party software would allow for the export of binary data to a text-based format, such as comma separated value files. Even in these situations, the data that would be needed for today's analysis might be strung over dozens, hundreds, or thousands of smaller files and each one would have to be exported by hand. (Note that GUI automation tools do exist for these type of tasks [in the academic world, this is where grad students would come in].)

To complicate matters further, this older software probably runs on an operating system that is just as old—often a flavor of Microsoft Windows. Thus, to convert the data into a more useable format, it may be necessary to spin up a virtual machine using this older operating system (but first you'll have to find a copy of the OS and a valid license!). In the worst case scenario, one might have to reverse-engineer the proprietary binary data format to unlock the data from the old silo. This process is often time consuming and could be of questionable legality. Reverse engineering requires a lot of significant extra time for data extraction. One new source of value that contemporary data explorations tap into is the ability to bring together multiple, disparate data sets. Thus, the above process may have to be replicated for each legacy data set that is brought into the new effort.

Let's assume however, that your data is not trapped in hundreds or thousands of proprietary binary files and is, instead, nestled inside a more familiar relational database. At first glance, one might assume that this scenario would offer smooth sailing, but this is not always the case. Relational databases a decade ago or older were far different beasts than they are today, and most were not open source. For example, PostgreSQL wasn't an open source database before 1996.[21] Thus, to stand up a duplicate copy of the database for analytic purposes you might need to acquire a copy and potentially even the appropriate license for the software, install the database, run the database, and maintain the database to a limited extent. At each stage of the process, you may face insurmountable obstacles—each one potentially completely denying access to the data that you need.

Contemporary Big Data Tools for the Traditional Engineer

For the traditional engineer who needs to get up to speed quickly on the evolution of big data, look no further than the flow of seminal papers from Google. From this stream you can see the big data challenges in the order that they arose and the technical approaches that Google used to address them. As Google has arguably been at the very front of the big data revolution, the sequence of innovation in the open source software world often mirrors Google's, just lagged by a few years.

In 2003, Google laid the most basic foundation for big data in terms of a distributed file system capable of handling truly big and unstructured data spread across thousands or millions of commodity machines with relative transparency to the end user.[22] They then provide a paradigm for processing that data at scale (MapReduce) in 2004, easing the cognitive burden of developers to increase productivity.[23] Next comes a way to handle structured data at scale (BigTable) in 2006 and then a system (Percolator) for incrementally

[21] Hellerstein, Joseph M., and Michael Stonebraker. Readings in Database Systems. Chapter 2, Cambridge, MA: MIT, 2005. Print.

[22] Ghemawat, Sanjay, Howard Gobioff, and Shun-Tak Leung. "The Google File System." *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles - SOSP '03* (2003). Print.

[23] Jeffrey Dean and Sanjay Ghemawat. 2004. "MapReduce: simplified data processing on large clusters." *Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation - Volume 6* (OSDI'04), Vol. 6. USENIX Association, Berkeley, CA, USA, 10-10.

updating their existing big data sets in 2010.[24,25] 2010 was a busy year for announcements as Google then addressed some of the shortfalls of MapReduce by telling the world about Pregel,[26] designed for large scale graph processing, and Dremel,[27] designed to handle near instantaneous interrogation of web-scale data, further increasing end-user productivity.

The big "G" returns to the world of relational databases, releasing two papers announcing new distributed systems that they have in production. First, in 2012, Google describes F1, a large scale distributed system that offers the scalability and fault tolerance of NOSQL databases and the transactional guarantees offered by a traditional relational database.[28] Second, in 2013, Google publicizes Spanner, a database distributed not just across cores and machines in a data center, but across machines distributed literally around the globe and the time synchronization problems encountered.[29] Last but certainly not least, Google discusses their dataflow model, an approach to processing data at scale that completely does away with the idea of a complete or finite data set required for batch processing. Instead, dataflow assumes that new data will always arrive and

[24] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber (2006), "Bigtable: A Distributed Storage System for Structured Data," Research (PDF), Google.

[25] Daniel Peng, and Frank Dabek. "Large-scale Incremental Processing Using Distributed Transactions and Notifications." *OSDI*. Vol. 10. 2010.

[26] Grzegorz Malewicz, et al. "Pregel: a system for large-scale graph processing." *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. ACM, 2010.

[27] Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, and Theo Vassilakis. 2010. "Dremel: interactive analysis of web-scale datasets." *Proc. VLDB Endow.* 3, 1-2 (September 2010), 330-339. DOI=10.14778/1920841.1920886

[28] Jeff Shute, et al. "F1: the fault-tolerant distributed RDBMS supporting google's ad business." *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, 2012.

[29] James C. Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, J. J. Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson Hsieh, Sebastian Kanthak, Eugene Kogan, Hongyi Li, Alexander Lloyd, Sergey Melnik, David Mwaura, David Nagle, Sean Quinlan, Rajesh Rao, Lindsay Rolig, Yasushi Saito, Michal Szymaniak, Christopher Taylor, Ruth Wang, and Dale Woodford. 2013. "Spanner: Google's Globally Distributed Database." *ACM Trans. Comput. Syst.* 31, 3, Article 8 (August 2013), 22 pages.

that old data may be retracted and that batch data processing is just a special case.[30]

Contemporary Data Storage

Data storage is the foundation upon which processing can occur and has evolved rapidly over the past two decades. Industrial scale databases are no longer dominated and controlled by proprietary commercial software from the Oracles of the world. Robust, scalable, and production-tested open source databases are available for free, one example being PostgreSQL. As relational databases aren't ideal for all types of data, a vast and somewhat confusing world of alternative datastores exist—document stores, graph, time series, in-memory, etc.—all suitable for handling a large variety of data and use cases, and all with pluses and minuses. Here, we will survey time series databases as they may be of significant interest to engineering-oriented companies.

Time Series Databases (TSDB)

In engineering, data is often generated by sensors and machines that produce new numeric values and associated timestamps at consistent, predetermined intervals. This is in stark contrast to much of the data seen in the Web 2.0, a world of social communication, messaging, and user interactions. In this world, data often comes in the form of actions performed by unpredictable humans at random time intervals. This fact helps explain why the time series database scene is significantly less evolved than that of the document store, which has already seen consolidation among market participants. If you need a NOSQL document store, MongoDB, RethinkDB, OrientDB, and others are happy to provide you with a different solutions. Likewise, if you are looking for a NOSQL datastore as a service, Amazon, Google, and many others provide numerous options.

However, TSDBs are now evolving quickly, partly due to the excitement around the Internet of Things. If sensors will be everywhere streaming measurements, we need data stores tailored to this partic-

[30] Tyler Akidau, et al. "The dataflow model: a practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing." *Proceedings of the VLDB Endowment* 8.12 (2015): 1792-1803.

ular use case. Another part of the driving force behind the advancement of TSDBs are the Googles and the Facebooks of the world. These companies have built their products and their businesses on the coordinated functioning of millions of servers. As these servers are continuously subjected to random hardware failures, these systems must be monitored. Even if we assume that we are only getting a few metrics per server per second, the amount of data adds up very quickly. For perspective, Facebook's TSDB, known as Gorilla, needed 1.3 terabytes of RAM to hold the last 26 hours of data in memory circa 2013.[31]

A time series database is designed from the ground up to handle time series data. What does this mean? First and foremost, TSDBs must always be available to accept and write time series data and, as we see from Facebook's example, the volume of data to be written can be extremely large. On the other side of the coin, read patterns are bursty and often produce aggregations (or roll ups) of the data over fixed windows. In terms of analytics, we often roll up time series into average or median values over certain periods (or windows) of time (a second, a minute, an hour, a day, etc.). For engineering problems, we may use the short-time Fourier transform on a windowed slice of data or get even more exotic using the Stockwell (S) transform.

The data that is getting stored is a sequence of numeric values coupled to time/date stamps plus associated metadata to describe the overall time series. There are creative ways to compress time stamps down to as small as a single bit per entry leveraging the consistent time interval at which they arrive and streaming numeric values that exploit temporal similarity in values and stores only the differences. Facebook claims to compress a single numeric value and corresponding time stamp, both 64-bit values, down to a total of 14 bits without loss of data.[32]

[31] Tuomas Pelkonen, Scott Franklin, Paul Cavallaro, Qi Huang, Justin Meza, Justin Teller, Kaushik Veeraraghavan, "Gorilla: A Fast, Scalable, In-Memory Time Series Database," *Proceedings of the VLDB Endowment*, Vol. 8, No. 12. 2015.

[32] *Ibid*

OpenTSDB

OpenTSDB is one of the more mature, open source time series databases and is currently at version 2.1.2. It was built in Java, designed to run on top of HBase as the backend data storage layer, and can handle millions of data points per second. OpenTSDB has been running in production for numerous large companies for the last 5-years.

InfluxDB, now InfluxData

InfluxData is a mostly open source, time series platform being built by a Series-A funded startup from New York City. Originally, the company was focused only on their time series database and experimented with multiple backend data storage engines before settling on their own in-house solution, the Time Structured Merge Tree. Now, InfluxData offers much of the functionality one would want for time series work in what they call the The TICK stack for time series data, composed of four different parts, mostly written in Go:

Telegraf
A data collection agent that helps collect time series data to be ingested into the database.[33]

InfluxDB
A scalable time series database designed to be dead simple to install.[34]

Chronograf
Time series data exploration tool and visualizer (not open source).

Kapacitor
Time series data processing framework for alerting and anomaly detection.[35]

InfluxData is still early (currently at release v0.10.1 as of February 2016) but has some large commercial partners and remains a promising option (until they are bought a la Titan?). This stack for working with time series makes a lot of sense as it addresses core needs of

[33] *https://github.com/influxdata/telegraf*

[34] *https://github.com/influxdata/influxdb*

[35] *https://github.com/influxdata/kapacitor*

users of time series data. However, one wonders if the component integration that InfluxData provides will prove more compelling than using best-of-breed alternatives built by third parties.

Cassandra

Apache Cassandra, originally developed at Facebook before being open sourced, is a massively scalable "database" that routinely handles petabytes of data in production for companies such as Apple and Netflix. While Cassandra was not designed for time series data specifically, a number of companies use it to store time series data. In fact, KairosDB is basically a fork of OpenTSDB that exchanges the original data storage layer, HBase, for Cassandra.

The core problem is that it requires *a lot* of extra developer time to realize much of the time series related functionality that you would want "built in." In fact, Paul Dix, the CEO and cofounder of InfluxData mentioned that InfluxDB arose from his experiences using Cassandra for time series work.

Processing Big Data

Once data sets expand past the size where a single machine can handle them, a distributed processing framework becomes necessary. One of the core conceptual differences between distributed computing frameworks is whether they handle data in batches or streaming (continuous). With batch processing, the data is assumed to be finite, regardless of size; it could be a yottabyte in size and spread across a million different servers. Hadoop is a batch processing framework and so is Spark to an extent as it uses microbatches. With streaming or unbounded data, we assume that the data will continue to arrive indefinitely, and thus, are working with an infinitely large data set. A lot of engineering data, including time series data, falls into the streaming or unbounded category. For example in the utility industry, synchrophasors (aka phasor measurement units or PMUs), report magnitude and angle measurements for every voltage and current phase up to 240 times per second. For a single line, this is 3 phases x 2 types x 2 x 240 = 2,880 samples per second for a single line.

If you are interested in a much deeper technical dive covering streaming versus batch processing, I cannot recommend the following two blog posts enough by Tyler Akidau at Google: Streaming 101 and Streaming 102.

From Hadoop to Spark (or from batch to microbatch)

The elephant in the room during any discussion of big data frameworks is always Hadoop. However, there is an heir apparent to the throne—Apache Spark—with more active developers in 2015 than any other big data software project. Spark came out of UC Berkeley's AMPLab (Algorithms, Machines, People) and became a top level Apache project in 2014. Even IBM has jumped on the bandwagon, announcing that it will put nearly 3,500 researchers to work using Spark moving forward.

Spark's meteoric rise to prominence can be explained by several factors. First, when possible, it keeps all data in memory, radically speeding up many types of calculations, including most machine learning algorithms that tend to be highly iterative in nature. This is in stark contrast to Hadoop that writes results to disk after each step. As disk access is much slower than RAM access, Spark can achieve 100x the performance over Hadoop for many machine learning applications.

Second, it comes equipped with a reasonably complete and growing toolkit for data. It's resilient distributed dataset provides a foundational data type for logically partitioning data across machines. SparkSQL allows simple connectivity to relational databases and a very useful dataframe object. GraphX offers tools for social network analysis and MLib does the same for machine learning. Finally, Spark Streaming helps to handle "real-time" data.

Third, it has done a great job courting the hearts and minds of data practitioners everywhere. While Java and Scala, Spark's native languages, aren't known for developer friendliness or rapid, iterative data exploration, Spark treats Python as a first class language and even plays well with IPython/Jupyter Notebook. This means practitioners can run their Python code on their own laptop using the same interface that they use to access a 1,000-node cluster. Speaking of a laptop, one of the most useful but poorly advertised features of Spark is the fact that just as it can leverage multiple cores across thousands of separate machines, it can do the same for a single laptop with multicore processor.

Next generation processing frameworks already?

Stream processing has made a big splash in the world of big data in 2015 and 2016. Fueling the need for streaming solutions has been the growing space of the Internet of Things and the industrial Internet of Things. Sensors will be connected to both consumer and industrial devices and these sensors will produce continuous updates for everything from your thermostat and light bulbs to the transformer outside your community.

To process this data, Google launched the Cloud Dataflow service —"a fully-managed cloud service and programming model for batch and streaming big data processing"—that is composed of two parts. The first part is the Cloud Dataflow SDKs that allow the end user to define the data and analysis needed for the job. Interestingly, these SDKs are becoming an Apache incubated project called Apache Beam. The second portion of the Cloud Dataflow service is the actual set of Google Cloud Platform technologies that allow the data analysis job to be run.

Alternatively, Apache Flink has emerged as an open source streaming data processing framework alternative to Google's Cloudflow service and as a potential competitor to Apache Spark. Apache Flink "is a streaming dataflow engine that provides data distribution, communication, and fault tolerance for distributed computations over data streams," that also has machine learning and graph processing libraries included by default. Originally, it was called Stratosphere and came out of a group of German universities including TU Berlin, Humboldt University, and the Hasso Plattner Institute; its first release as an Apache project was in August of 2014. Now, there are at least two options to process streaming data at scale using either Google's cloud based offering or building out your own system with Apache Flink.

Geomagnetic Disturbances—A Case Study of Approaches

Geomagnetic Disturbances (GMDs here on out) represent a significant stochastic threat to the power grid of the United States of America. They also present an interesting case study to compare and contrast traditional engineering, data science, and even hybrid approaches to tackling what has been a challenging problem for the industry.

A Little Space Science Background

To start, let's provide a little background. The Earth has a magnetic field that emanates from the flow of molten charged metals in its core. This geomagnetic field extends into space and, as with any magnet, has both a north and a south pole. Geomagnetic North and South are not the same as the North and South Poles but they are reasonably close.

Our star, a swirling ball of superheated plasma, ejects vast clouds of charged particles at high speed from across its surface. This solar wind is composed mostly of protons and electrons traveling around 450–700 kilometers per second. This wind is occasionally interrupted by coronal mass ejections, violent eruptions of plasma from the sun at different trajectories. These trajectories sometimes intersect the Earth's orbit and, occasionally, the Earth itself with a glancing blow or a direct hit. These charged particles interact with the Earth's magnetosphere and ionosphere with several consequences.

Most beautifully and benignly, charged particles from the sun can actually enter Earth's atmosphere, directed to the North and South Magnetic poles by the magnetosphere. Once in the atmosphere, the charged particles excite atoms of atmospheric gases such as nitrogen and oxygen. To relax back to their normal state, these atoms emit the colorful lights that we refer to as the northern lights or the aurora borealis in the northern hemisphere. In the southern hemisphere, this phenomenon is called the southern lights or the aurora australis.[36]

Unfortunately, the auroras are not the only effect. Charged particles arising from coronal mass ejections interacting with the magnetosphere can temporarily distort and perturb the Earth's magnetic field known as geomagnetic disturbances (GMDs). A time varying magnetic field can induce large currents in the power grid called geomagnetically induced currents (GICs). GICs are considered quasi-DC currents because they oscillate far slower than the 60 Hz frequency of alternating current used by the North American grid. GICs flow along high voltage transmission lines and then go to ground through high voltage transformers. Having large amplitude direct current flowing through a transformer can cause half cycle

[36] *http://www.noaa.gov/features/monitoring_0209/auroras.html*

saturation, generating harmonics in the power system and heating the windings of the transformer. While these issues might not sound too bad, unchecked heating can destroy the transformer and sufficient harmonics can trigger failsafe devices, bringing down parts or all of the grid. GICs have also been linked to audible noise described in some cases as if the transformer were growling.[37]

Questioning Assumptions

When we take a closer look at the GMD phenomenon, we find some interesting assumptions present in the industry that may or may not be accurate. Despite a vast amount of research into our magnetosphere, there is much left to discover in terms of the interactions with Earth. For example, recent research utilizing high performance computing to create a global simulation of the Earth-ionosphere waveguide under the effect of a geomagnetic storm,[38] has exposed a previously unknown coupling mechanism between coronal mass ejections and the Earth's magnetosphere. In other words, even our best physics-based models do not yet fully explain the behavior that we have witnessed.

Geomagnetically induced currents are often associated with high voltage equipment and this is where a bulk of the research is focused. Higher voltage lines have lower resistances and thus experience larger GICs. Further, higher voltage transformers are more expensive and take much longer to repair or replace and are thus of more interest to study. However, there is at least statistical evidence that GMDs impact equipment and businesses consuming power at the other end of the power grid. More specifically, Schrijver, et al. examined over eleven thousand insurance claims during the first decade of the new millennium and found that claim rates were elevated by approximately 20% on days in the top 5% of elevated geomagnetic activity.[39] Further, the study suggests "that large-scale geomagnetic variability couples into the low-voltage power distribu-

[37] "Effects of Geomagnetic Disturbances on the Bulk Power System," February 2012, North American Electric Reliability Corporation.

[38] Jamesina Simpson, University of Utah. "Petascale Computing: Calculating the Impact of a Geomagnetic Storm on Electric Power Grids."

[39] C. J. Schrijver, R. Dobbins, W. Murtagh, and S. M. Petrinec. "Assessing the Impact of Space Weather on the Electric Power Grid Based on Insurance Claims for Industrial Electrical Equipment." *Space Weather* 12.7 (2014): 487-98. Print.

tion network and that related power-quality variations can cause malfunctions and failures in electrical and electronic devices that, in turn, lead to an estimated 500 claims per average year within North America."

GMDs have always been associated with far northern (or southern) latitudes that are closer to the magnetic poles. Interestingly, there is new evidence that interplanetary shocks can cause equatorial geomagnetic disturbances whose magnitude is enhanced by the equatorial electrojet.[40] This is very noteworthy for at least two reasons. First, such shock waves may or may not occur during what is traditionally thought of as a geomagnetic storm. Thus, a GMD could occur during a "quiet period" with literally no warning. Second, this phenomenon impacts utilities and power equipment closer to the equator, a region where components of the power grid are not thought to need GMD protection.

The impact of GMDs and GICs, while not completely instantaneous, have always been assumed to be immediate and not long term in nature. However, Gaunt and Coetzee found that GICs may impact power grids lying between 18 and 30 degrees South that were traditionally thought to be at low risk. Second, and potentially more importantly, it would appear that small geomagnetically induced currents may be capable of creating longer term damage to transformers that reduces the lifespan of the equipment, causing equipment failures that occur months after a GMD.[41]

Solutions

The seemingly high impact, low frequency (HILF) nature of geomagnetic disturbances has presented problems for the industry and the industry's regulatory bodies. Let's suppose for a moment that, unlike contemporary thinking, GICs are a near omnipresent, low-level occurrence. How this strain manifests in large transformers over extended exposure is unknown and likely random in nature; small inhomogeneities in materials unknown during the manufac-

[40] B. A. Carter, E. Yizengaw, R. Pradipta, A. J. Halford, R. Norman, and K. Zhang. "Interplanetary Shocks and the Resulting Geomagnetically Induced Currents at the Equator." *Geophys. Res. Lett. Geophysical Research Letters* 42.16 (2015): 6554-559. Print.

[41] C. T. Gaunt, and G. Coetzee. "Transformer Failures in Regions Incorrectly Considered to Have Low GIC-risk." *2007 IEEE Lausanne Power Tech* (2007). Print.

ture of components cause uneven stresses and strains that aren't captured by contemporary physics-based models. On the other end of the severity spectrum, how does one prepare for the 50-year or even 100-year storm, similar to the 1859 Carrington Event, that could offer near apocalyptic consequences for the country and even society. The stochastic nature of this insult to the grid is part of the core problem of devising and implementing solutions.

The traditional engineering approach

The traditional engineering approach attacks the problem leveraging the known physics underlying GIC current flows. If the resistance increases along the path to ground through the transformer, the current will flow somewhere else. Currently, there are smart devices on the market that act as metallic grounds for transformers but, in the presence of GIC flows, interrupt the ground, replacing it with both a series resistor and capacitor to block currents up to a specified threshold. While this can protect a particular transformer, the current will still flow to ground somewhere, potentially impacting a different part of the system. Further, there is an obvious and large capital equipment expense purchasing and installing a separate device for each transformer to be protected.

Extending the engineering approach—the Hydro One real-time GMD management system

Canada, due to its northern latitude and direct experiences with GMD, has been at the forefront of GMD research and potential solutions. It is only fitting that Hydro One in Toronto is the first utility with a real-time GMD management system in operation. This system, almost by necessity, combines the traditional engineering approaches standardized in the industry—physics-based models that are updated periodically with coarse grain measurements—with new sensors operated by the utility and an external data source driving additional modeling efforts.

In more detail, the Hydro One SCADA system collects voltage measurements on the grid and power flow through transformers as is a common practice of utilities. More impressive and much less standard, Hydro One also measures GIC neutral currents from 18 stations, harmonics from 5 transformers, dissolved gas analysis telemetry from 6 monitors, and transformer and station ambient temperature. Further, the magnetometer in Ottawa run by the Canadian

Magnetic Observatory System (CANMOS) supplies 1 Hz magnetic field measurements batch updated each minute. This magnetic field data is then combined with previous measurements of ground conductivity in the region to compute the geoelectric field value. The resulting geoelectric field then drives a numerical model that computes GIC flows throughout the system.[42] Where GICs are not being directly monitored by a physical sensor, they are being computed with a model that can be verified continuously. Thus, Hydro One has, in essence, extended the traditional engineering-based approach with the integration of near real time data to address the GMD issue.

The purely data driven detection approach

Over the last decade, the Department of Energy has helped utilities deploy nearly two thousand synchrophasors or PMUs to take real-time, high fidelity sensor measurements of the grid. The current SCADA system captures measurements once every few seconds. However, PMUs measure the current and voltage phasors anywhere from 15 to 240 times per second, several orders of magnitude faster than the current SCADA system.

If one has an accurate record of when transformers on the grid have experienced geomagnetically induced currents, this record can be used as ground truth. This ground truth can be associated via timestamps to the historical PMU data to create a labeled training set. With this labeled training set, any number of supervised learning approaches could be used and then validated to build a potential GIC detector.

The purely data-driven predictive approach

One potential purely data driven approach would be to steal a page from the Panopticon's playbook and leverage a very broad data set to attempt to predict imminent geomagnetic disturbances. With sufficient lead time and low enough false alarm rates, utilities could take preventative steps to mitigate the impact of GMDs on the power grid with warning.

[42] Luis Marti, and Cynthia Yin. "Real-Time Management of Geomagnetic Disturbances: Hydro One's Extreme Space Weather Control Room Tools." *IEEE Electrification Magazine IEEE Electrific. Mag.* 3.4 (2015): 46-51. Print.

Such a diverse and potentially predictive data set exists across a number of government agencies. The USGS runs the Geomagnetism program that operates 14 observatories streaming sensor measurements of the Earth's magnetic field. Adding to this pool of measurements is the Canadian Magnetic Observatory System with 14 additional magnetic observatories in North America (see Figure 1-4). While 28 magnetometer sensors don't nearly cover the entire North American continent, they do provide some insight into the immediate behavior of the geomagnetic field. Further, as GMDs tend to be multihour and even multiday events, intraevent structure could allow for a predictive warning even just from real-time magnetometer data.

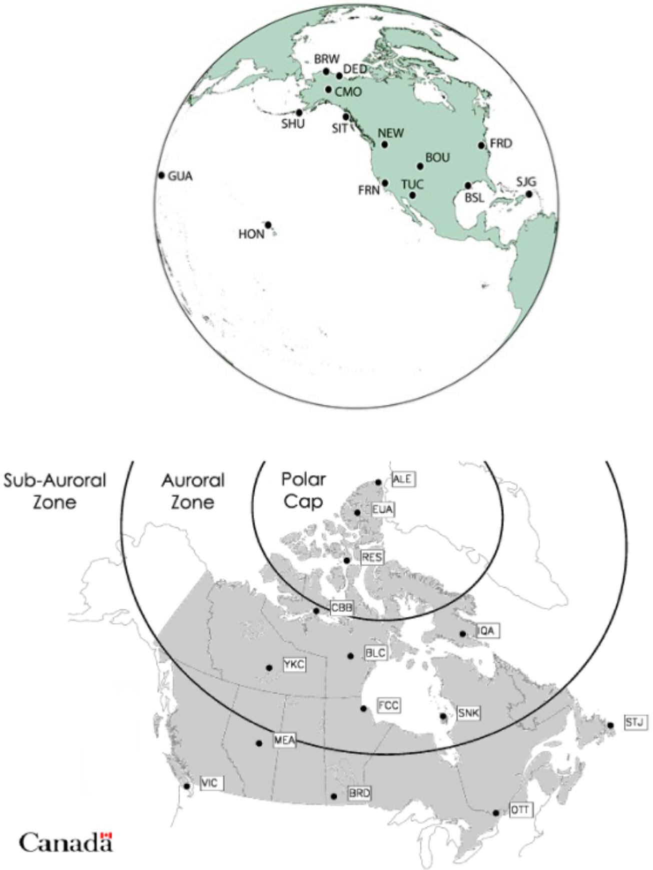

Figure 1-4. Magnetic observatories in North America

If more lead time is needed, multiple space-based satellites are equipped with sensors that provide potentially relevant data. The Geostationary Operational Environmental Satellites (GOES) sit in geosynchronous orbit and many have operational magnetometers. At this altitude, the GOES satellites potentially offer up to 90 seconds of warning about potential geomagnetic disturbances.

If even more lead time is needed, NOAA's Deep Space Climate Observatory (DSCOVR) is set to replace the ACE (Advanced Composition Explorer), both in stable orbits between the Earth and sun at the Lagrange point L1. DSCOVR can measure solar wind speeds and other aspects of space weather, providing warnings at least 20 minutes in advance of an actual event. Taken together, it is possible that these data streams could support the prediction of accurate warnings of GMD events on Earth.

Conclusion

The above are only a small sampling of the approaches that could be taken to address geomagnetic disturbances and it is clear that the use of data will factor heavily into most options. PingThings is currently working on what could be considered a hybrid approach to this problem. We are using high data rate sensors combined with a physics-based understanding of the grid's operation to bring quantified awareness of GIC to the power grid at a cost significantly lower than hardware-based strategies.

More broadly, there are many more challenges that the nation's grid faces with everything from squirrels to cyberterrorists threatening to turn off the lights. As the electric utilities are not the only engineering-based companies that find themselves facing such issues, data science and machine learning will continue to infiltrate existing legacy industries. While these deterministic models and machines have always existed in our stochastic world, we now have the tools and techniques to better address this reality; the evolution is inevitable.

About the Author

Sean Patrick Murphy serves as the Chief Data Scientist for PingThings, an Industrial Internet of Things (IIoT) startup bringing advanced data science and machine learning to the nation's electric grid. He is a founder and board member of Data Community DC, a 10,000-member community of data practitioners, and leads the 1,500+ member Data Innovation DC MeetUp that focuses on the use of data for value creation.

He completed his graduate work in biomedical engineering at Johns Hopkins University and stayed on as a senior scientist at the Johns Hopkins University Applied Physics Laboratory for over a decade where he focused on machine learning, anomaly detection, image analysis, and high performance and cloud-based computing. He graduated from the DC inaugural class of the Founder Institute, completed Hacker School in New York City, and serves as a judge and mentor for Venture for America.